SPE e a Comunidade

Análise de Sobrevivência em R

Ivo Sousa-Ferreira, *ivo.ferreira@staff.uma.pt*

Departamento de Estatística e Investigação Operacional, Faculdade de Ciências, Universidade de Lisboa, Portugal e CEAUL – Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Portugal
e
Departamento de Matemática, Faculdade de Ciências Exatas e da Engenharia, Universidade da Madeira, Portugal

A Análise de Sobrevivência é um dos ramos mais antigos da Estatística, com origem no século XVII. Surgiu, essencialmente, como resposta à necessidade de obter métodos estatísticos que permitissem a resolução de problemas na área das Ciências Biomédicas, o que influenciou toda a terminologia utilizada. O seu foco é a análise do tempo de vida, que é definido como o tempo que decorre desde um instante inicial até à ocorrência de um acontecimento de interesse. Este acontecimento é estipulado à partida e pode assumir diversas formas como, por exemplo, a recidiva de uma doença, a avaria de uma máquina, a união matrimonial, a conclusão de um ciclo de estudos ou a dissolução de uma sociedade comercial. Assim, consoante a natureza do acontecimento em estudo, este ramo tem sido aplicado nas mais variadas áreas, nomeadamente na Medicina, Engenharia, Ciências da Educação, Sociologia e Economia. Adicionalmente, a escolha deste tipo de análise estatística tem a particularidade de permitir a inclusão de dados censurados, que surgem quando, por algum motivo, não é possível observar o acontecimento de interesse durante o período em que os indivíduos estão em observação. Para um conhecimento mais aprofundado, consulte-se Lawless (2003).

Os métodos tradicionais de Análise de Sobrevivência encontram-se disponíveis em vários programas de *software* estatístico, sendo que alguns dos mais utilizados são: SPSS, SAS, STATA, S-Plus e R. No entanto, a maioria desses programas é também bastante rudimentar no que diz respeito à análise de dados de sobrevivência, pois oferece pouca diversidade metodológica ao utilizador. Em oposição, o *software* estatístico R destaca-se por ser uma ferramenta bastante abrangente, uma vez que fornece uma panóplia vastíssima de métodos que procuram solucionar os mais variados problemas de Estatística e Probabilidade. Além disso, este programa beneficia do facto de ser gratuito, o que se revela uma mais-valia para toda a comunidade académica, em especial na área da investigação, permitindo que qualquer cientista ou organização possa contribuir para o seu desenvolvimento.

O *software* estatístico R, doravante designado apenas por R, foi concebido no início da década de 90 do século XX, pelos estatísticos Ross Ihaka e Robert Gentleman da Universidade de Auckland, em Nova Zelândia. Desde 1997, tem sido mantido por uma equipa, denominada por R Core Team (2022), a qual é constituída por 20 cientistas e programadores. O seu *download* de acesso livre pode ser feito através do sítio eletrónico da rede de arquivos do R (CRAN – *Comprehensive R Archive Network*). De entre as diferentes características que apresenta, o R possui um vasto leque de pacotes, em constante crescimento e atualização, o que lhe confere potencialidades acrescidas à sua já poderosa versão base. Porém, a abundância de pacotes (à data do presente texto, encontravam-se disponíveis mais de 19000 pacotes) pode ser vista como um obstáculo na procura rápida e eficiente do(s) pacote(s) mais adequado(s) para um determinado problema. Para ultrapassar essa questão, foram desenvolvidas *CRAN task views* sobre tópicos específicos (atualmente, existem 40 *task views*), que visam fornecer orientações para encontrar o pacote mais relevante para executar uma dada tarefa (*task*).

A Análise de Sobrevivência usufrui da vantagem de ter uma *task view* exclusivamente dedicada a si. A *task view Survival* é da autoria de Allignol e Latouche (2022), que também têm a responsabilidade de efetuar a sua atualização/manutenção regular. Esta *task view* está organizada em 12 temas centrais: Análise de Sobrevivência *standard*; modelos multiestados; sobrevivência relativa, modelos de efeitos aleatórios, sobrevivência multivariada, modelos bayesianos; aprendizagem automática; predições e desempenho da predição; análise de potência; simulação; gráficos; e miscelânea. Ao longo destes temas, existe um pacote que é referido com frequência por ser muito generalista no que toca às ferramentas que disponibiliza ao utilizador sobre uma grande variedade de temas – o pacote **survival** (Therneau, 2021).

O pacote **survival**, assim como outros pacotes do R, dispõe de um conjunto de vinhetas que tem como objetivo fornecer tutoriais instrutivos para exemplificar a utilização prática das suas funcionalidades e auxiliar na interpretação e discussão de resultados. A vinheta *A package for survival analysis in R* é aquela que, no meu ponto de vista, providencia uma excelente introdução sobre este pacote e aborda mais assuntos relacionados com a análise de dados de sobrevivência. Nessa vinheta, encontra-se uma breve história a respeito da sua origem, onde é referido que as suas funções começaram por ser desenvolvidas para o S-Plus, em 1985. Como ambos os programas, R e S-Plus, têm por base a linguagem de programação S, muitos dos códigos desenvolvidos para o S-Plus podem ser facilmente implementados no R e, por vezes, nem é necessário efetuar qualquer alteração a esses códigos. Deste modo, a transição do S-Plus para o R ocorreu de forma natural, sem o autor indicar a data em que tal aconteceu.

Presentemente, o pacote **survival** é um dos pacotes mais utilizados no estudo da variável tempo de vida e, de acordo com o autor da vinheta supracitada, uma das principais razões para o sucesso deste pacote é o facto de todas as suas funções terem sido escritas para solucionar problemas associados à análise de dados reais, não descurando a fundamentação teórica que sustenta e atribui significado às funções programadas. Como é evidente, o referido pacote não cobre todos os temas deste ramo da estatística, mas seguramente fornece as rotinas e os objetos basilares para a definição de funções mais complexas. Uma prova disso é a existência de mais de 800 pacotes (dependentes) cujas funções são definidas à custa das funções do pacote **survival**. Além do mais, este é um dos poucos pacotes *core* do R, pacotes esses que são instalados em conjunto, e de forma automática, com a versão base do *software*.

Na literatura, é possível encontrar alguns livros que ilustram o modo como o pacote **survival** pode ser usado para modelar dados de sobrevivência e estimar certas quantidades de interesse, como sejam a probabilidade de sobrevivência e o risco relativo. De seguida, passo a indicar 3 livros que, no meu entender, enriquecem o leitor com uma excelente visão geral sobre os princípios fundamentais da Análise de Sobrevivência e as suas aplicações através do R. Um dos livros foi redigido por Therneau e Grambsch (2000) e, ao longo dos seus exemplos práticos, apresenta código do S-Plus que pode ser facilmente implementado no R com recurso ao pacote **survival** (Therneau, 2021). Os outros dois livros são da autoria de Carvalho *et al.* (2011) e de Moore (2016) que, embora abordem maioritariamente as particularidades associadas aos comandos do pacote **survival**, também apresentam código do R referente ao uso de outros pacotes como, por exemplo, o pacote **mstate** (de Wreede *et al*., 2011) para estimar a função de incidência cumulativa em cenários de riscos competitivos; e o pacote **muhaz** (Gentleman, 2021) para efetuar a estimação da função de risco usando métodos não paramétricos baseados em funções *kernel*. Importa referir que, de entre estes três livros, apenas a obra de Carvalho *et al.* (2011) está escrita em português, aspeto este que pode ser encorajador para os principiantes em R que tenham preferência pela língua portuguesa.

O livro de Moore (2016) faz parte da série de livros *Use R!*, publicada pela editora Springer. Conforme descrito por esta editora, os livros da referida série destinam-se a apresentar uma discussão acerca do uso do R numa área específica (*e.g.*, biometria, econometria e quimiometria) ou num tópico estatístico em particular (*e.g.*, dados omissos, dados espaciais e dados longitudinais). A série *Use R!* tem no mínimo mais 2 livros que se enquadram no domínio da Análise de Sobrevivência, nomeadamente as obras de Beyersmann *et al.* (2012) e de Willekens (2014). Ambos os livros exemplificam a utilização de diversos pacotes do R para ajustar modelos multiestado, ainda que o primeiro livro dê especial atenção à modelação de causas competitivas. Existem muitas outras editoras bem conceituadas que publicam obras

relacionadas com o R, sendo que algumas dessas obras estão reportadas na página principal deste *software*, mais propriamente na secção de documentação (consulte-se *Books related to R*).

Outra forma autodidata de manusear este programa, consiste em aceder a artigos de revistas científicas centradas no desenvolvimento de investigação em Estatística Computacional. Desde 2009, o próprio R tem uma revista indexada de acesso aberto, cujos artigos são sujeitos a revisão por pares. Esta revista, denominada *The R Journal*, prioriza a publicação de artigos que abordam aspetos computacionais relacionados com as novas ferramentas ou pacotes do R.

Apesar do R permitir a programação de qualquer função através da sua interface de linha de comando, a sua versão base não incorpora uma interface gráfica do utilizador (GUI) que promova uma interação amigável, por meio de elementos gráficos. Consequentemente, aprender a manusear este programa é, sem dúvida, uma tarefa árdua para os iniciantes em programação. Para ultrapassar esse obstáculo, têm sido desenvolvidas GUI para o R. O *R Commander* é, provavelmente, a GUI mais antiga do R, tendo sido construída no formato de um pacote, designado por **Rcmdr** (Fox, 2005), o que significa que a sua instalação e carregamento são feitos da maneira habitual. Esta interface possibilita a implementação de funções através de um simples clique no seu menu de opções, além de continuar a permitir a escrita de linhas de comando.

Associado ao **Rcmdr** existem ainda os *plugins*, que são, na realidade, pacotes do R que visam estender as funcionalidades dos menus do *R Commander*. Neste momento, existem mais de 40 *plugins* para o **Rcmdr** e alguns deles são bastante úteis no âmbito da Análise de Sobrevivência, onde considero relevante destacar o **RcmdrPlugin.survival** (Fox e Carvalho, 2012) e o **RcmdrPlugin.EZR**. (Kanda, 2013). O primeiro *plugin* acrescenta algumas das funcionalidades do pacote **survival** aos menus do *R Commander*, incluindo o modelo de Cox e os modelos de sobrevivência paramétricos. Segundo Carvalho *et al.* (2011), o **RcmdrPlugin.survival** surgiu da necessidade em apresentar, de uma forma gentil, as potencialidades do pacote **survival** aos médicos investigadores. Relativamente ao segundo *plugin*, este também foi especialmente criado para facilitar a análise de dados em saúde. Note-se que, em inglês, a sigla EZR é lida como "*easy R"*. Em comparação com o *plugin* anterior, o **RcmdrPlugin.EZR** trata uma maior variedade de temas, incluindo a análise da curva ROC (*Receiver Operating Characteristic*) associada a um modelo de sobrevivência, a meta-análise do risco relativo, o cálculo da dimensão da amostra, entre outros.

Para uma breve revisão sobre outros pacotes do R com elevada utilidade neste ramo da estatística, remete-se o leitor para o trabalho de Abreu e Gouveia-Reis (2016), publicado nas atas do XXII Congresso da SPE. Nesse artigo, as autoras referem que a dinâmica do R faz com que qualquer trabalho sobre este *software* seja inevitavelmente incompleto e um pouco desatualizado. Esta é uma realidade com a qual os utilizadores do R têm que lidar, a qual está bem espelhada na evolução do seu número de pacotes, considerando que à data da escrita desse artigo (setembro de 2016) existiam cerca de 9000 pacotes e, em apenas 5 anos e meio, passaram a existir mais do dobro!

Por fim, não poderia deixar de realçar que existem vários investigadores portugueses, distribuídos ao longo de todo o território nacional (sem esquecer os arquipélagos atlânticos), que têm vindo a contribuir de algum modo para o desenvolvimento de pacotes do R no ramo da Análise de Sobrevivência, alguns em colaboração com cientistas de outros países. Na Tabela 1 estão indicados alguns desses pacotes por ordem alfabética, sem diminuir o mérito de outros trabalhos que mereciam igualmente ser aqui reportados.

Agradecimentos

O presente texto foi redigido na vigência da bolsa de doutoramento DFA/BD/6459/2020 e do projeto UIDB/00006/2020 (CEAUL – Centro de Estatística e Aplicações), financiados por Fundos Nacionais através da FCT – Fundação para a Ciência e a Tecnologia.

Tabela 1: Alguns pacotes do R para a análise de dados de sobrevivência, desenvolvidos por (ou com a colaboração de) investigadores portugueses

Designação	Referência	Breve descrição
clustcurv	Villanueva *et al.* (2021)	Um método para determinar *clusters* em curvas múltiplas com seleção automática do seu número, tendo por base os algoritmos *k*-médias ou *k*-medianas. A metodologia pode ser aplicada para determinar *clusters* em várias curvas de sobrevivência.
condSURV	Meira-Machado e Sestelo (2016)	Fornece alguns estimadores não paramétricos e semiparamétricos (os quais são extensões do estimador de Kaplan-Meier) para estimar a função de sobrevivência condicional, na análise multivariada de tempos por indivíduo.
genSurv	Araújo *et al.* (2021)	Simulação de dados de sobrevivência com uma covariável dependente do tempo e simulação de dados de sobrevivência a partir de um modelo progressivo de doença-morte.
JMbayes2	Rizopoulos *et al.* (2021)	Permite realizar o ajustamento de modelos conjuntos para dados longitudinais e de sobrevivência, sob a abordagem bayesiana.
joineR	Philipson *et al.* (2021)	Análise de medições repetidas e de dados de sobrevivência por meio da sua modelação conjunta com efeitos aleatórios. Os dados de sobrevivência são modelados usando o modelo de Cox com covariáveis dependentes do tempo, enquanto o *outcome* longitudinal é modelado usando um modelo linear de efeitos mistos. A associação é capturada através de um processo gaussiano latente.
smoothHR	Meira-Machado *et al.* (2013)	Permite o cálculo das estimativas pontuais do risco relativo (e correspondentes intervalos de confiança) associado a uma covariável contínua, considerando que esta tem um efeito não linear que é definido à custa de uma função *spline* penalizada.
survidm	Soutinho *et al.* (2021)	Apresenta alguns métodos desenvolvidos recentemente para estimar várias quantidades associadas ao modelo de doença-morte, tais como a probabilidade de transição, a probabilidade de ocupação, a função de incidência cumulativa e a distribuição do tempo de permanência em cada estado.
vsd	Camacho e Abreu (2021)	Visualização de vários gráficos para dados de sobrevivência com censura à direita, incluindo a representação da estimativa de Kaplan-Meier da função de sobrevivência, da estimativa suavizada da função de risco e dos gráficos de floresta referentes aos coeficientes de regressão de um modelo de sobrevivência.

Referências

Abreu, A. M. e Gouveia-Reis, D. (2016). Análise de sobrevivência e valores extremos em R. Em C. R. Cordeiro, *Estatística: Progressos e Aplicações. Atas do XXII Congresso da Sociedade Portuguesa de Estatística* (pp. 1 - 14). Portugal: Edições SPE.

Allignol, A. e Latouche, A. (2022). *CRAN Task View: Survival Analysis*. Consultado a 8 de fevereiro de 2022. URL: https://cran.r-project.org/web/views/Survival.html

Araújo, A., Meira-Machado, L. e Faria, S. (2021). *genSurv: generating multi-state survival data*. Pacote do R versão 1.0.4. URL: https://CRAN.R-project.org/package=genSurv

Beyersmann, J., Allignol, A. e Schumacher, M. (2012). *Competing Risks and Multistate Models with R*. New York: Springer.

Camacho, D. e Abreu, A. M. . (2021). *vsd: graphical shim for visual survival data analysis*. Pacote do R versão 0.1.0. URL: https://CRAN.R-project.org/package=vsd

Carvalho, M. S., Andreozzi, V. L., Codeço, C. T., Campos, D. P., Barbosa, M. T. S. e Shimakura, S. E. (2011). *Análise de Sobrevivência: Teoria e Aplicações em Saúde* (2.ª ed.). Rio de Janeiro, Brasil: Fiocruz.

de Wreede, L. C., Fiocco, M. e Putter, H. (2011). mstate: an R package for the analysis of competing risks and multi-state models. *Journal of Statistical Software, 38*(7), 1-30.
Fox, J. (2005). The R Commander: a basic-statistics graphical user interface to R. *Journal of Statistical Software, 19*(4), 1-42.
Fox, J. e Carvalho, M. S. (2012). The RcmdrPlugin.survival package: extending the R Commander interface to survival analysis. *Journal of Statistical Software, 49*(7), 1-32.
Gentleman, R. (2021). *muhaz: hazard function estimation in survival analysis*. Pacote do R versão 1.2.6.4. URL: https://CRAN.R-project.org/package=muhaz
Kanda, Y. (2013). Investigation of the freely available easy-to-use software 'EZR' for medical statistics. *Bone Marrow Transplant, 48*, 452–458.
Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data* (2.ª ed.). New York: John Wiley & Sons.
Meira-Machado, L. e Sestelo, M. (2016). condSURV: an R package for the estimation of the conditional survival function for ordered multivariate failure time data. *The R Journal, 8*(2), 460-473.
Meira-Machado, L., Cadarso-Suárez, C., Gude, F. e Araújo, A. (2013). smoothHR: an R package for pointwise nonparametric estimation of hazard ratio curves of continuous predictors. *Computational and Mathematical Methods in Medicine*, 1-11.
Moore, D. F. (2016). *Applied Survival Analysis Using R*. New York: Springer.
Philipson, P., Sousa, I., Diggle, P., Williamson, P., Kolamunnage-Dona, R., Henderson, R. e Hickey, G. (2021). *joineR: joint modelling of repeated measurements and time-to-event data*. Pacote do R versão 1.2.6. URL: https://CRAN.R-project.org/package=joineR
R Core Team (2022). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. Vienna, Austria. URL: https://www.R-project.org/
Rizopoulos, D., Papageorgiou, G. e Afonso, P. M. (2021). *JMbayes2: extended joint models for longitudinal and time-to-event data*. Pacote do R versão 0.1-8. URL: https://CRAN.R-project.org/package=JMbayes2
Soutinho, G., Sestelo, M. e Meira-Machado, L. (2021). survidm: an R package for inference and prediction in an illness-death model. *The R Journal*, 1-20.
Therneau, T. M. (2021). *survival: A package for survival analysis in S*. Pacote do R versão 3.2-13: https://CRAN.R-project.org/package=survival
Therneau, T. M. e Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. New York: Springer-Verlag.
Villanueva, N. M., Sestelo, M., Meira-Machado, L. e Roca-Pardiñas, J. (2021). clustcurv: an R package for determining groups in multiple curves. *The R Journal, 13*(1), 164-183.
Willekens, F. (2014). *Multistate Analysis of Life Histories with R*. New York: Springer.